MATRIX CONCENTRATION FOR PRODUCTS

DE HUANG, JONATHAN NILES-WEED, JOEL A. TROPP, AND RACHEL WARD

Abstract. This paper develops nonasymptotic growth and concentration bounds for a product of independent random matrices. These results sharpen and generalize recent work of Henriksen–Ward, and they are similar in spirit to the results of Ahlswede–Winter and of Tropp for a sum of independent random matrices. The argument relies on the uniform smoothness properties of the Schatten trace classes.

1. Motivation

Products of random matrices arise in many contemporary applications in the mathematics of data science. For instance, they describe the evolution of stochastic linear dynamical systems, which include popular stochastic algorithms for optimization such as Oja's algorithm for streaming principal component analysis [28] and the randomized Kaczmarz method for solving linear systems [36]. To understand the detailed behavior of these algorithms, such as the rate of convergence, we may seek out methods for studying a product of random matrices.

Unfortunately, the tools currently available in the literature are poorly adapted to these circumstances. Indeed, an instantiation of a stochastic optimization algorithm involves a finite product of finite-dimensional matrices, often with a particular structure (e.g., low-rank perturbations of the identity). But most existing theoretical results are limit laws that require the number of factors in the product or the dimension of the factors to tend to infinity. Furthermore, strong assumptions on the random matrices (e.g., independent and identically distributed entries) are usually required.

This paper offers some new tools for studying random matrix products that arise from stochastic optimization algorithms and related problems. The research is inspired by the recent paper [19] of Henriksen and Ward. Our hope is to replicate the successful program for studying sums of random matrices, implemented in the works [1, 29, 38, 39, 40, 41]. In particular, we seek to develop methods that are flexible, easy to use, and powerful [42]. We also aspire to use transparent theoretical arguments that can be adapted easily to new situations.

2. Contributions

To motivate our work, we start with an elementary concentration inequality for a product of independent random numbers. We will generalize this bound, and others, to the matrix setting.

2.1. Context: A Product of Random Numbers Near 1.

Consider an independent family $\{X_1, X_2, \dots\} \subset \mathbb{R}$ of bounded random variables that satisfy

$$\mathbb{E}\, X_i = \mu \quad \text{and} \quad |X_i - \mu|^2 \le b^2 \quad \text{almost surely.}$$

Date: 4 March 2020.

The authors gratefully acknowledge the funding for this work. DH was supported under NSF grant DMS-1613861. JNW and RW were supported in part by the Institute for Advanced Study, where some of this research was conducted. JAT was supported under ONR Awards N00014-17-1-2146 and N00014-18-1-2363. RW also received support from AFOSR MURI Award N00014-17-S-F006.

Form a product of random perturbations of 1, and compute its mean:

$$Z_n := \prod_{i=1}^n \left(1 + \frac{X_i}{n}\right) \quad \text{and} \quad \mathbb{E}\, Z_n = \left(1 + \frac{\mu}{n}\right)^n = \mathrm{e}^{\mu} \cdot (1 - O(n^{-1})).$$

We anticipate that the random product Z_n concentrates around its expectation $\mathbb{E}\, Z_n \approx \mathrm{e}^{\mu}$.

To check this surmise, we can use standard methods from scalar concentration theory. For $s > 0$,

$$\begin{aligned}
\mathbb{P}\left\{Z_n \geq (1+s)\,\mathrm{e}^{\mu}\right\} &= \mathbb{P}\left\{\prod_{i=1}^n \left(1 + \frac{X_i}{n}\right) \geq (1+s)\,\mathrm{e}^{\mu}\right\} \\
&\leq \mathbb{P}\left\{\exp\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \geq (1+s)\,\mathrm{e}^{\mu}\right\} \\
&= \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}\, X_i) \geq \log(1+s)\right\}.
\end{aligned}$$

The inequality follows from the numerical fact $1 + a \leq \mathrm{e}^a$, valid for $a \in \mathbb{R}$. Hoeffding's inequality furnishes the bound

$$\mathbb{P}\left\{Z_n \geq (1+s)\,\mathrm{e}^{\mu}\right\} \leq \exp\left(\frac{-n\log^2(1+s)}{2b^2}\right). \tag{2.1}$$

At the small scale $s \leq \mathrm{e}$, in which case $\log(1+s) \geq s/\mathrm{e}$, the growth bound (2.1) implies a subgaussian tail behavior:

$$\mathbb{P}\left\{Z_n - \mathbb{E}\, Z_n \geq t\,\mathrm{e}^{\mu}\right\} \leq \mathbb{P}\left\{Z_n - \mathrm{e}^{\mu} \geq t\,\mathrm{e}^{\mu}\right\} \leq \exp\left(\frac{-nt^2}{2\mathrm{e}^2 b^2}\right) \quad \text{for } t \leq \mathrm{e}. \tag{2.2}$$

A similar inequality holds for the lower tail.

2.2. **A Product of Random Perturbations of the Identity.** We might hope that products of random matrices exhibit a similar behavior. Consider an independent family $\{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\} \subset \mathbb{M}_d$ of $d \times d$ matrices that satisfy

$$\mathbb{E}\, \boldsymbol{X}_i = \boldsymbol{A} \quad \text{and} \quad \|\boldsymbol{X}_i - \mathbb{E}\, \boldsymbol{X}_i\|^2 \leq b^2 \quad \text{almost surely.} \tag{2.3}$$

Here are elsewhere, $\|\cdot\|$ is the spectral norm, that is, the ℓ_2 operator norm. Form a product of random perturbations of the identity and compute its mean:

$$\boldsymbol{Z}_n = \left(\mathbf{I} + \frac{\boldsymbol{X}_n}{n}\right) \cdots \left(\mathbf{I} + \frac{\boldsymbol{X}_1}{n}\right) \quad \text{and} \quad \mathbb{E}\, \boldsymbol{Z}_n = \left(\mathbf{I} + \frac{\boldsymbol{A}}{n}\right)^n \approx \mathrm{e}^{\boldsymbol{A}}. \tag{2.4}$$

Is it true that the spectral norm $\|\boldsymbol{Z}_n\|$ is proportional to e^{μ}, where $\mu = \|\boldsymbol{A}\|$? Does the random product $\boldsymbol{Z}_n$ concentrate near its mean $\mathbb{E}\, \boldsymbol{Z}_n$?

These speculations are correct. Moreover, we can obtain bounds that parallel the scalar inequalities announced in the last subsection. Here is one particular result that follows from our analysis.

Theorem I (Products of Perturbations of the Identity—Special case)**.** *Consider an independent family $\{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\} \subset \mathbb{M}_d$ of random matrices that satisfy the hypotheses (2.3). Define $\mu := \|\boldsymbol{A}\|$. The matrix product $\boldsymbol{Z}_n$ introduced in (2.4) satisfies the bounds*

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n\| \geq (1+s)\,\mathrm{e}^{\mu}\right\} \leq d \cdot \exp\left(\frac{-n\log^2(1+s)}{2b^2}\right) \qquad \textit{when } \log(1+s) \geq 2b^2/n;$$

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n - \mathbb{E}\, \boldsymbol{Z}_n\| \geq t\mathrm{e}^{\mu}\right\} \leq (d+\mathrm{e}) \cdot \exp\left(\frac{-nt^2}{2\mathrm{e}^2 b^2}\right) \qquad \textit{when } t \leq \mathrm{e}.$$

Theorem I follows from Corollary 6.1.

As compared with the scalar bounds (2.1) and (2.2), the results in Theorem I feature an additional dimensional factor d in front of the exponential. This term leads to a dependency of $\log d$ in the bounds for products of random matrices. Otherwise, everything is the same, including the constants.

2.3. **Proof Strategy.** How might one establish a result like Theorem I? The derivation in Section 2.1 is valid only for products of random scalars. We cannot even begin to make this argument for matrices because the exponential of a sum of matrices generally does not equal the product of the exponentials.

In this paper, we take a completely different approach. The key is to observe that multiplying a random product $\boldsymbol{Z} \in \mathbb{M}_d$ by a statistically independent factor $\boldsymbol{Y} \in \mathbb{M}_d$ creates a predictable change plus a random perturbation:

$$\boldsymbol{YZ} = (\mathbb{E}\,\boldsymbol{Y})\boldsymbol{Z} + (\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y})\boldsymbol{Z}.$$

Since the second term has zero mean, conditional on $\boldsymbol{Z}$, we can exploit this orthogonality property to estimate the size of the product:

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{YZ}\|_2^2 &= \mathbb{E}\,\|(\mathbb{E}\,\boldsymbol{Y})\boldsymbol{Z}\|_2^2 + \mathbb{E}\,\|(\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y})\boldsymbol{Z}\|_2^2 \\ &\le \left(\|\mathbb{E}\,\boldsymbol{Y}\|^2 + \mathbb{E}\,\|\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}\|^2\right)\left(\mathbb{E}\,\|\boldsymbol{Z}\|_2^2\right) =: (1+v)\,m \cdot \left(\mathbb{E}\,\|\boldsymbol{Z}\|_2^2\right)\end{aligned}$$

The notation $\|\cdot\|_2$ refers to the Schatten 2-norm, also known as the Frobenius norm. The last step introduces data about the random matrix $\boldsymbol{Y}$: the mean $m = \|\mathbb{E}\,\boldsymbol{Y}\|$ and the relative variance $v = \mathbb{E}\,\|\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}\|^2 / \|\mathbb{E}\,\boldsymbol{Y}\|^2$. We can apply the same argument recursively to decompose the matrix $\boldsymbol{Z}$ into its own factors.

The approach in the last paragraph depends on the fact that $\|\cdot\|_2$ is the norm induced by the trace inner product. To undertake the same action for the spectral norm $\|\cdot\|$, we first need to approximate the spectral norm by the Schatten p-norm for $p \approx \log d$. Then we can invoke a remarkable geometric property of the Schatten p-norm, called *uniform smoothness*, as a substitute for the orthogonality law. See the paper [26] for an introduction to this circle of ideas. Section 4 executes this method.

2.4. **Additional Results.** We establish a family of norm inequalities for products of random matrices. The main result, Theorem 5.1, gives a bound for the moments of a Schatten p-norm of a random product and a centered random product. From this fact, we derive expectation bounds, tail bounds, and matrix concentration inequalities. Many of these results hold under weaker assumptions than Theorem I, addressing cases where the matrices have different means or are unbounded.

To give a better indication of what we can prove, let us give an informal presentation of one of our main results, Corollary 5.4. The statement concerns a general product $\boldsymbol{Z}_n = \boldsymbol{Y}_n \cdots \boldsymbol{Y}_1$ of independent random matrices of dimension d. Abbreviating $p = 1 + 2\log d$, we have the inequality

$$\mathbb{E}\,\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \le \mathrm{e}\sqrt{pv}\,\prod\nolimits_{i=1}^{n} \|\mathbb{E}\,\boldsymbol{Y}_i\| \quad \text{when} \quad v := \sum\nolimits_{i=1}^{n} \frac{\mathbb{E}\,\|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i\|^2}{\|\mathbb{E}\,\boldsymbol{Y}_i\|^2} \le \frac{1}{p}.$$

We can interpret v as the accumulated relative variance in the product.

For example, in the setting of Theorem I, the quantity $v = O(b^2/n)$. It follows that

$$\mathbb{E}\,\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| = O\left(\sqrt{\frac{pb^2}{n}}\,\|\mathbb{E}\,\boldsymbol{Z}_n\|\right). \tag{2.5}$$

In particular, $\|\boldsymbol{Z}_n\|$ is much closer to e^{μ} than to the worst-case bound e^{b}.

2.5. **Roadmap.** We continue with an overview of related work in Section 3. Section 4 presents background results from matrix theory and high-dimensional probability. We establish our main results for general matrix products in Section 5. Afterward, Section 6 draws corollaries for a product of perturbations of the identity. Finally, we describe some refinements and extensions in Section 7.

3. Related Work

Products of random matrices have been studied for decades, primarily within the fields of ergodic theory, control theory, random matrix theory, and free probability. More recently, applied mathematicians have developed results that are tailored to problems arising in data science. Almost all prior work is either asymptotic in the length of the product or asymptotic in the dimension of the matrices. This section contains an overview of these inquiries.

3.1. **Direct Connections.** The most immediate precedent for our research is the recent paper of Henriksen and Ward [19]. They were motivated by the problem of understanding streaming algorithms for covariance estimation. Their work gives, perhaps, the first explicit nonasymptotic bounds for a somewhat general product of random matrices with fixed dimension. The argument is based on the matrix Bernstein inequality and a combinatorial fact about set partitions.

Henriksen and Ward focus on the setting of Theorem I, and they establish a bound of the form

$$\mathbb{E}\,\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \le \frac{b\mathrm{e}^b}{\sqrt{n}} \cdot \operatorname{polylog}(n, d).$$

In contrast, our new result (2.5) replaces the worst-case factor e^b with the more typical value e^μ. We are also able to relax several of the assumptions in [19].

Also in the setting of Theorem I, several works obtain results on the asymptotic behavior of $\boldsymbol{Z}_n$. Berger [8] establishes, via a semigroup argument based on the Chernoff product formula, that $\boldsymbol{Z}_n \to \mathrm{e}^{\boldsymbol{A}}$ in probability as $n \to \infty$. Emme and Hubert [13] recently obtained a refinement of this result: motivated by a problem in ergodic theory, they show that $\boldsymbol{Z}_n \to \mathrm{e}^{\boldsymbol{A}}$ as $n \to \infty$ under the sole assumptions that $\sum_{i=1}^n \boldsymbol{X}_i/n \to \boldsymbol{A}$ and $\sum_{i=1}^n \|\boldsymbol{X}_i\|/n < \infty$. Their argument expands the product and computes the limit of the kth order term using an induction. Neither approach readily yields nonasymptotic bounds.

3.2. **Other Recent Applications.** Some applied work on random matrix products has been driven by the empirical observation that stochastic gradient descent converges faster when the gradient approximations are sampled *without* replacement, rather than sampled *with* replacement. Some papers that investigate this question from the point of view of (nonasymptotic) matrix inequalities include [32, 20, 2]. This specific problem has been solved by Gürbüzbalaban et al. [17] using optimization theory. However, none of these results directly address the questions at hand.

Researchers studying randomly initialized deep neural networks have also developed theoretical analysis for products of random matrices; see [18, 46]. These results involve operations on matrices with independent entries, and they focus on the large-matrix limit.

3.3. **Ergodic Theory and Control Theory.** Products of random matrices describe the evolution of a linear stochastic dynamical system. For this reason, they have been a subject of perennial interest within the literatures on ergodic theory and on control theory. For the most part, this research is concerned with properties of the asymptotics of infinite products of matrices (of fixed size). Let us give a few more details.

Consider a finite family $\mathscr{A} = \{\boldsymbol{A}_1, \dots, \boldsymbol{A}_s\} \subset \mathbb{M}_d$ of fixed matrices. Construct a random matrix $\boldsymbol{X} \in \mathbb{M}_d$ with the distribution

$$\mathbb{P}\,\{\boldsymbol{X} = \boldsymbol{A}_i\} = \frac{1}{s} \quad \text{for each } i = 1, \dots, s.$$

The *Lyapunov exponent* of the set $\mathscr{A}$ is the quantity

$$\lambda(\mathscr{A}) := \lim_{n\to\infty} \frac{1}{n} \log \|\boldsymbol{X}_n \cdots \boldsymbol{X}_1\| \quad \text{where } \boldsymbol{X}_i \sim \boldsymbol{X} \text{ iid.}$$

The Furstenberg–Kesten theorem [15] establishes that $\lambda(\mathscr{A})$ exists almost surely, but approximating $\lambda(\mathscr{A})$ is algorithmically undecidable [43, Thm. 2]. As a consequence, we must be pessimistic about finding a completely satisfactory solution to the matrix concentration problem for products.

To learn more about Lyapunov exponents and to find additional references, see the paper [3] for work in control theory and the paper [45] for work in ergodic theory. Another major application of random products is to study the asymptotic behavior of a random walk on a group; we refer the reader to [23, 14, 7] for more information.

3.4. **Random Matrix Theory and Free Probability.** Products of random matrices have also been considered within random matrix theory and free probability. This connection is natural, but matrix products have received somewhat less attention than other kinds of random matrix models. In these contexts, it is common to study a product of a small number of matrices (two or three, say) in the limit as the dimension of the matrices grows.

Bai and Silverstein [4, Chap. 4] present a limit law for the sequence of products of a random matrix with iid entries and a random matrix whose spectral distribution has a deterministic limit. This theorem is motivated by a statistical application, multivariate analysis of variance. Note, however, that convergence of the spectral distribution does not determine the limit of the spectral norm.

Free probability gives a complete description of the spectral distribution of a product of two freely independent elements as the "multiplicative free convolution" of the spectral distributions of the factors. The connection to random matrix theory stems from the fact that a family of "adequately random" matrices becomes freely independent in the limit as the dimension of the matrices tends to infinity. See the book of Nica & Speicher [27] for a digestible introduction; some other good treatments include [31, 34, 35]. Free probability has significant applications in wireless communications [44].

For highly structured random matrices (invariant ensembles), it may be possible to obtain more detailed formulas for products. See [21, 12] for some recent work in this direction.

4. Random Matrix Inequalities via Uniform Smoothness

To analyze products of random matrices, we exploit classic methods that were developed to study the evolution of a martingale taking values in a uniformly smooth Banach space. These ideas are relevant for us because the matrix Schatten classes (with power $2 \le p < \infty$) enjoy a remarkable uniform smoothness property.

In this section, we outline the required background from matrix analysis and high-dimensional probability. Naor's tutorial paper [26] serves as a model for our presentation, and it contains a more general treatment. See Section 4.6 for additional discussion about the history of these ideas.

4.1. **Notation and Background.** We work in the complex field $\mathbb{C}$; identical results hold for the real field $\mathbb{R}$. We often use the infix notation for the minimum ($\wedge$) and the maximum ($\vee$) of two real numbers.

The operator $\mathbb{P}$ computes the probability on an event. The operator $\mathbb{E}$ computes the expectation of a random variable. Subscripts denote partial expectation; for example, $\mathbb{E}_Z$ is the expectation over the randomness in Z. Nonlinear functions, such as powers, bind before the expectation.

The linear space $\mathbb{C}^{d\times r}$ contains all $d \times r$ matrices with complex entries. The algebra $\mathbb{M}_d$ consists of all $d \times d$ matrices with complex entries. We use the standard definitions of scalar multiplication, matrix addition, matrix multiplication, and the adjoint (i.e., conjugate transpose). Any statement about matrices that is not qualified with specific dimensions holds for all matrices with compatible dimensions. Nonlinear functions, such as matrix powers, bind before the trace. The matrix absolute value $|\boldsymbol{A}| := (\boldsymbol{A}^*\boldsymbol{A})^{1/2}$, where $(\cdot)^{1/2}$ is the positive-semidefinite square root of a positive-semidefinite matrix.

We write $\|\cdot\|$ for the spectral norm on matrices; the spectral norm coincides with the maximum singular value, and it is also known as the ℓ_2 operator norm. For each $p \geq 1$, the symbol $\|\cdot\|_p$ refers to the Schatten p-norm which returns the ℓ_p norm of the singular values of its argument. The symbol S_p refers to a linear space of matrices (of fixed dimension), equipped with the Schatten p-norm.

For parameters $p, q \geq 1$, we define the $L_q(S_p)$ norm of a random matrix $\boldsymbol{X}$ as

$$\|\!|\!|\boldsymbol{X}|\!|\!|_{p,q} := \|\boldsymbol{X}\|_{L_q(S_p)} := \left(\mathbb{E}\,\|\boldsymbol{X}\|_p^q\right)^{1/q}.$$

The $L_q(S_p)$ norm is an operator ideal norm, in the sense that

$$\|\!|\!|\boldsymbol{A}\boldsymbol{X}|\!|\!|_{p,q} \leq \|\boldsymbol{A}\| \cdot \|\!|\!|\boldsymbol{X}|\!|\!|_{p,q} \quad \text{for fixed } \boldsymbol{A} \text{ and random } \boldsymbol{X}. \tag{4.1}$$

This statement follows instantly from the analogous property of the Schatten p-norm.

We sometimes use the following simple inequalities for the moments of a random matrix $\boldsymbol{X}$:

$$\mathbb{E}\,\|\boldsymbol{X}\| \leq \inf_{p\geq 1} \mathbb{E}\,\|\boldsymbol{X}\|_p = \inf_{p,q\geq 1} \|\!|\!|\boldsymbol{X}|\!|\!|_{p,q}. \tag{4.2}$$

The equality follows from Lyapunov's inequality, combined with the fact that $\|\!|\!|\boldsymbol{X}|\!|\!|_{p,1} = \mathbb{E}\,\|\boldsymbol{X}\|_p$ for all $p \geq 1$.

4.2. **Uniform Smoothness for Matrices.** Uniform smoothness [1] is a property of a normed space that describes how much the norm of a point changes under symmetric perturbation. Since the Schatten-2 space S_2 is an inner-product space, the parallelogram law gives an exact description of this phenomenon:

$$\frac{1}{2}\left[\|\boldsymbol{X}+\boldsymbol{Y}\|_2^2 + \|\boldsymbol{X}-\boldsymbol{Y}\|_2^2\right] = \|\boldsymbol{X}\|_2^2 + \|\boldsymbol{Y}\|_2^2.$$

Remarkably, in other Schatten classes, the parallelogram law is replaced by an inequality.

Fact 4.1 (Uniform Smoothness for Schatten Classes)**.** *Let $\boldsymbol{A}, \boldsymbol{B}$ be matrices of the same size. For $p \geq 2$,*

$$\left[\frac{1}{2}\left(\|\boldsymbol{A}+\boldsymbol{B}\|_p^p + \|\boldsymbol{A}-\boldsymbol{B}\|_p^p\right)\right]^{2/p} \leq \|\boldsymbol{A}\|_p^2 + \mathrm{C}_p\,\|\boldsymbol{B}\|_p^2. \tag{4.3}$$

The optimal constant $\mathrm{C}_p := p-1$. The inequality is reversed when $1 \leq p \leq 2$.

Fact 4.1 was first established by Tomczak-Jaegermann [37]; she obtained the sharp constant C_p when p is an even number. Ball, Carlen, and Lieb [5, Thm. 1] determined that C_p is the optimal constant for all values of p. Throughout the paper, we will continue to write $\mathrm{C}_p = p-1$.

4.3. **Uniform Smoothness for Random Matrices.** Much as the Schatten class S_p of matrices enjoys a uniform smoothness property, the normed space $L_q(S_p)$ of random matrices is also uniformly smooth. When $2 \leq q \leq p$, this statement follows as an easy consequence of Fact 4.1.

Corollary 4.2 (Uniform Smoothness for Random Matrices)**.** *Let $\boldsymbol{X}, \boldsymbol{Y}$ be random matrices of the same size. When $2 \leq q \leq p$,*

$$\left[\frac{1}{2}\left(\|\!|\!|\boldsymbol{X}+\boldsymbol{Y}|\!|\!|_{p,q}^q + \|\!|\!|\boldsymbol{X}-\boldsymbol{Y}|\!|\!|_{p,q}^q\right)\right]^{2/q} \leq \|\!|\!|\boldsymbol{X}|\!|\!|_{p,q}^2 + \mathrm{C}_p\|\!|\!|\boldsymbol{Y}|\!|\!|_{p,q}^2.$$

Proof. Apply Lyapunov's inequality to the left-hand side of (4.3) to pass from the pth power to the qth power, and then transfer the exponent to the right-hand side to obtain the pointwise bound

$$\frac{1}{2}\left(\|\boldsymbol{X}+\boldsymbol{Y}\|_p^q + \|\boldsymbol{X}-\boldsymbol{Y}\|_p^q\right) \leq \left[\|\boldsymbol{X}\|_p^2 + \mathrm{C}_p\,\|\boldsymbol{Y}\|_p^2\right]^{q/2}.$$

[1] More precisely, we are considering uniformly smooth spaces whose modulus of smoothness has power type 2.

Take the expectation, and use the triangle inequality for the $L_{q/2}$ norm:

$$\frac{1}{2}\left(\mathbb{E}\,\|\boldsymbol{X}+\boldsymbol{Y}\|_p^q + \mathbb{E}\,\|\boldsymbol{X}-\boldsymbol{Y}\|_p^q\right) \le \left[\left(\mathbb{E}\,\|\boldsymbol{X}\|_p^q\right)^{2/q} + \mathrm{C}_p\left(\mathbb{E}\,\|\boldsymbol{Y}\|_p^q\right)^{2/q}\right]^{q/2}.$$

Reinterpret the latter display using the $L_q(S_p)$ norm $|||\cdot|||_{p,q}$. □

4.4. **Subquadratic Averages for Random Matrices.** Corollary 4.2 admits a powerful extension that controls how the norm of a matrix changes if we add a random matrix that has zero mean. This result is the main tool that we employ in our study of random products.

Proposition 4.3 (Subquadratic Averages)**.** *Consider random matrices $\boldsymbol{X}, \boldsymbol{Y}$ of the same size that satisfy $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}] = \mathbf{0}$. When $2 \le q \le p$,*

$$|||\boldsymbol{X}+\boldsymbol{Y}|||_{p,q}^2 \le |||\boldsymbol{X}|||_{p,q}^2 + \mathrm{C}_p|||\boldsymbol{Y}|||_{p,q}^2.$$

The constant $\mathrm{C}_p = p-1$ is the best possible.

Ricard and Xu [33] obtained a version of Proposition 4.3 in the more general setting of a von Neumann algebra. In their work, the expectation implicit in the L_q norm is replaced by the projection onto a subalgebra. They emphasize that the key feature of their work is the determination of the sharp constant.

Here, we offer a very short proof of Proposition 4.3 with a suboptimal constant. The method is drawn from Naor's paper [26]. Lemma A.1, in the appendix, unspools an elementary argument that delivers the sharp constant.

Proof. By Jensen's inequality, applied conditionally on $\boldsymbol{X}$,

$$\begin{aligned}\frac{1}{2}\left(|||\boldsymbol{X}+\boldsymbol{Y}|||_{p,q}^2 + |||\boldsymbol{X}|||_{p,q}^2\right) &\le \frac{1}{2}\left(|||\boldsymbol{X}+\boldsymbol{Y}|||_{p,q}^2 + |||\boldsymbol{X}-\boldsymbol{Y}|||_{p,q}^2\right)\\ &\le \left[\frac{1}{2}\left(|||\boldsymbol{X}+\boldsymbol{Y}|||_{p,q}^q + |||\boldsymbol{X}-\boldsymbol{Y}|||_{p,q}^q\right)\right]^{2/q} \le |||\boldsymbol{X}|||_{p,q}^2 + \mathrm{C}_p|||\boldsymbol{Y}|||_{p,q}^2.\end{aligned}$$

The second inequality is Lyapunov's; the third is Corollary 4.2. Upon rearranging, we find that

$$|||\boldsymbol{X}+\boldsymbol{Y}|||_{p,q}^2 \le |||\boldsymbol{X}|||_{p,q}^2 + 2\mathrm{C}_p|||\boldsymbol{Y}|||_{p,q}^2. \tag{4.4}$$

This is the stated result, with a spurious factor of 2. □

4.5. **Matrix-Valued Martingales.** To demonstrate the value of Proposition 4.3, let us explain how it leads to moment bounds for a matrix-valued martingale sequence. Consider a null matrix martingale $\{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\} \subset \mathbb{M}_d$ with difference sequence $\{\boldsymbol{\Delta}_1, \dots, \boldsymbol{\Delta}_n\} \subset \mathbb{M}_d$. That is,

$$\boldsymbol{X}_0 = \mathbf{0} \quad \text{and} \quad \boldsymbol{X}_i = \boldsymbol{X}_{i-1} + \boldsymbol{\Delta}_i \quad \text{where} \quad \mathbb{E}[\boldsymbol{\Delta}_i|\boldsymbol{X}_0, \dots, \boldsymbol{X}_{i-1}] = \mathbf{0} \quad \text{for } i = 1, \dots, n.$$

Applying Proposition 4.3 repeatedly, we arrive at the bound

$$|||\boldsymbol{X}_n|||_{p,q}^2 \le \mathrm{C}_p \sum\nolimits_{i=1}^n |||\boldsymbol{\Delta}_i|||_{p,q}^2. \tag{4.5}$$

In words, the squared norm of the martingale is controlled by the sum of the squares of the norms of the martingale differences. The inequality (4.5) is a powerful extension of the orthogonality of the increments of a martingale taking values in an inner-product space, say S_2. The uniform smoothness constant C_p shows how the geometry of the matrix space intermediates.

In this work, we will develop bounds for random matrix products by applying a similar technique to appropriately chosen decompositions of the product.

4.6. **History.** The approach in this section has a long history. Let us summarize the contributions that are most relevant to our development.

For real numbers, the (sharp) uniform smoothness property in Fact 4.1 is known as the *two-point inequality*; it was established independently by Leonard Gross [16] and Aline Bonami [10] in the early 1970s, with later contributions by William Beckner [6]. In 1974, the uniform smoothness property for the Schatten classes was obtained by Nicole Tomczak-Jaegermann [37]. It took another 20 years before Ball, Carlen, and Lieb [5] obtained the sharp uniform smoothness constants for all Schatten classes. The property dual to uniform smoothness is called *uniform convexity*. See [5] for a detailed exposition.

Tomczak-Jaegermann [37, Thm. 3.1] also demonstrated that Rademacher averages are subquadratic in each Schatten space S_p with $p \geq 2$; that is, the Banach space S_p is *type 2* [22]. This fact is a prototype for the more general result stated in Proposition 4.3. Tropp [39, Sec. 4.8] points out that parts of the Ahlswede–Winter [1, App.] theory of sums of independent random matrices already follow from Tomczak-Jaegermann's work. (In contrast, Tropp's matrix concentration inequalities [39] are more closely related to a fact from operator theory, the noncommutative Khintchine inequality of Françoise Lust-Piquard [25]; Tropp's results are derived using a theorem [24, Thm. 6] of Elliot Lieb.)

Assaf Naor [26] traces the application of uniform convexity inequalities in the study of martingales to a 1975 paper of Gilles Pisier [30]. Naor [26] gives a nice introduction to this circle of ideas, which he uses to derive a general version of the Azuma inequality that holds in any uniformly smooth Banach space.

At least as early as 1988, Donald Burkholder [11] applied closely related convexity inequalities to derive sharp inequalities for martingales taking values in a Hilbert space. The paper [33] of Éric Ricard and Quanhua Xu is a recent entry in this line of research.

5. A Product of Independent Random Matrices

In this section, we obtain our main results on the growth and concentration of a product of independent random matrices. Section 5.1 shows how to decompose a random product into pieces that we can control using a recursive argument. Based on these ideas, we derive Theorem 5.1, a general bound on the moments of the norm of the matrix product. The moment estimate leads to a family of expectation bounds (Corollary 5.4) and probability bounds (Corollary 5.6).

The balance of the paper contains applications of these results (Section 6) and extensions of the method to other settings (Section 7).

5.1. **Decomposition of Random Products.** Our approach is based on a recursive argument that describes how the product evolves as we include more factors. At each step, we decompose the product into a nonrandom term and a random term with mean zero. This formulation allows us to apply Proposition 4.3 on subquadratic averages.

Consider a fixed matrix $\boldsymbol{Z}_0 \in \mathbb{M}_d$ and an independent family $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_n\} \subset \mathbb{M}_d$ of random matrices. We can recursively construct products of these random matrices:

$$\boldsymbol{Z}_i = \boldsymbol{Y}_i \boldsymbol{Z}_{i-1} \quad \text{for } i = 1, \dots, n.$$

Evidently, the last element of the sequence takes the form $\boldsymbol{Z}_n = \boldsymbol{Y}_n \cdots \boldsymbol{Y}_1 \boldsymbol{Z}_0$. By independence, $\mathbb{E}\,\boldsymbol{Z}_n = (\mathbb{E}\,\boldsymbol{Y}_n) \cdots (\mathbb{E}\,\boldsymbol{Y}_1)\boldsymbol{Z}_0$.

The random product $\boldsymbol{Z}_i$ admits a simple decomposition into a mean term and a fluctuation term:

$$\boldsymbol{Z}_i = \boldsymbol{Y}_i \boldsymbol{Z}_{i-1} = (\mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{Z}_{i-1} + (\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{Z}_{i-1} \quad \text{for each } i = 1, \dots, n. \tag{5.1}$$

Since $\boldsymbol{Y}_i$ is independent from $\boldsymbol{Z}_{i-1}$, the second term is conditionally zero mean:

$$\mathbb{E}[(\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{Z}_{i-1} | \boldsymbol{Z}_{i-1}] = \boldsymbol{0}. \tag{5.2}$$

The property (5.2) supports the use of Proposition 4.3. It is also helpful to have an explicit norm bound for the random fluctuation term:

$$|||(\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{Z}_{i-1}|||_{p,q} \le \left(\mathbb{E}\,\|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i\|^q \cdot \mathbb{E}\,\|\boldsymbol{Z}_{i-1}\|_p^q\right)^{1/q} = \left(\mathbb{E}\,\|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i\|^q\right)^{1/q} |||\boldsymbol{Z}_i|||_{p,q}. \tag{5.3}$$

The first relation follows from the operator ideal property of the Schatten p-norm and the statistical independence of the random matrices $\boldsymbol{Y}_i$ and $\boldsymbol{Z}_{i-1}$.

We can study the concentration properties of the product $\boldsymbol{Z}_i$ using a related decomposition:

$$\boldsymbol{Z}_i - \mathbb{E}\,\boldsymbol{Z}_i = \boldsymbol{Y}_i\boldsymbol{Z}_{i-1} - (\mathbb{E}\,\boldsymbol{Y}_i)(\mathbb{E}\,\boldsymbol{Z}_{i-1}) = (\mathbb{E}\,\boldsymbol{Y}_i)(\boldsymbol{Z}_{i-1} - \mathbb{E}\,\boldsymbol{Z}_{i-1}) + (\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{Z}_{i-1}. \tag{5.4}$$

As in (5.2), the second term is a fluctuation that is conditionally zero mean. The fluctuation term satisfies the norm bound (5.3).

5.2. **Growth and Concentration.** Our main result controls the growth of the moments of a product of independent random matrices. It also describes how well the random product concentrates around its expectation.

Theorem 5.1 (Growth and Concentration of Products)**.** *Consider a fixed matrix $\boldsymbol{Z}_0 \in \mathbb{C}^{d\times r}$ and an independent family $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_n\} \subset \mathbb{M}_d$ of random matrices. Form the product*

$$\boldsymbol{Z}_n = \boldsymbol{Y}_n\boldsymbol{Y}_{n-1}\cdots\boldsymbol{Y}_2\boldsymbol{Y}_1\boldsymbol{Z}_0 \in \mathbb{C}^{d\times r}.$$

For parameters $2 \le q \le p$, assume that

$$\|\mathbb{E}\,\boldsymbol{Y}_i\| \le m_i \quad \text{and} \quad \left(\mathbb{E}\,\|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i\|^q\right)^{1/q} \le \sigma_i m_i \quad \text{for } i = 1, \dots, n.$$

Define the product of means and the accumulated relative variance

$$M = \prod\nolimits_{i=1}^n m_i \quad \text{and} \quad v = \sum\nolimits_{i=1}^n \sigma_i^2.$$

Then the random product $\boldsymbol{Z}_n$ satisfies the growth bound and the concentration bound

$$|||\boldsymbol{Z}_n|||_{p,q} \le \mathrm{e}^{\mathrm{C}_p v/2}\,\|\boldsymbol{Z}_0\|_p \cdot M; \tag{5.5}$$

$$|||\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n|||_{p,q} \le \left(\mathrm{e}^{\mathrm{C}_p v} - 1\right)^{1/2} \|\boldsymbol{Z}_0\|_p \cdot M. \tag{5.6}$$

Proof of Theorem 5.1, relation (5.5). By the homogeneity of (5.5), we may assume that $m_i = 1$ for each index i, so that also $M = 1$. As in (5.1), we have the decomposition

$$\boldsymbol{Z}_i := \boldsymbol{Y}_i\boldsymbol{Z}_{i-1} = (\mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{Z}_{i-1} + (\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{Z}_{i-1} \quad \text{for each } i = 1, \dots, n.$$

Now, Proposition 4.3 implies that

$$\begin{aligned}
|||\boldsymbol{Z}_i|||_{p,q}^2 &\le |||(\mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{Z}_{i-1}|||_{p,q}^2 + \mathrm{C}_p \cdot |||(\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{Z}_{i-1}|||_{p,q}^2 \\
&\le \|\mathbb{E}\,\boldsymbol{Y}_i\|^2 \cdot |||\boldsymbol{Z}_{i-1}|||_{p,q}^2 + \mathrm{C}_p\left(\mathbb{E}\,\|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i\|^q\right)^{2/q} \cdot |||\boldsymbol{Z}_{i-1}|||_{p,q}^2 \\
&\le (1 + \mathrm{C}_p\sigma_i^2) \cdot |||\boldsymbol{Z}_{i-1}|||_{p,q}^2 \\
&\le \exp(\mathrm{C}_p\sigma_i^2) \cdot |||\boldsymbol{Z}_{i-1}|||_{p,q}^2.
\end{aligned}$$

The second line follows from (5.3), and the third depends on our hypotheses about the factors $\boldsymbol{Y}_i$. The last relation requires the numerical inequality $1 + a \le \mathrm{e}^a$, valid for all $a \in \mathbb{R}$. By iteration,

$$|||\boldsymbol{Z}_i|||_{p,q}^2 \le \exp\left(\mathrm{C}_p \sum\nolimits_{k=1}^i \sigma_k^2\right) \cdot \|\boldsymbol{Z}_0\|_p^2. \tag{5.7}$$

In the final step, we use the assumption that $\boldsymbol{Z}_0$ is not random to see that $|||\boldsymbol{Z}_0|||_{p,q} = \|\boldsymbol{Z}_0\|_p$. For $i = n$, the formula (5.7) is the advertised result. □

Proof of Theorem 5.1, relation (5.6). The pattern of argument is similar with the proof of (5.5). By the homogeneity of (5.6), we may assume that all $m_i = 1$ and that $M = 1$. As in (5.4), we have the

decomposition

$$\boldsymbol{Z}_i - \mathbb{E}\,\boldsymbol{Z}_i = \boldsymbol{Y}_i\boldsymbol{Z}_{i-1} - (\mathbb{E}\,\boldsymbol{Y}_i)(\mathbb{E}\,\boldsymbol{Z}_{i-1}) = (\mathbb{E}\,\boldsymbol{Y}_i)(\boldsymbol{Z}_{i-1} - \mathbb{E}\,\boldsymbol{Z}_{i-1}) + (\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{Z}_{i-1}.$$

Again, we invoke Proposition 4.3 to ascertain that

$$\begin{aligned}
|||\boldsymbol{Z}_i - \mathbb{E}\,\boldsymbol{Z}_i|||_{p,q}^2 &\le |||(\mathbb{E}\,\boldsymbol{Y}_i)(\boldsymbol{Z}_{i-1} - \mathbb{E}\,\boldsymbol{Z}_{i-1})|||_{p,q}^2 + \mathrm{C}_p \cdot |||(\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{Z}_{i-1}|||_{p,q}^2 \\
&\le |||\boldsymbol{Z}_{i-1} - \mathbb{E}\,\boldsymbol{Z}_{i-1}|||_{p,q}^2 + \mathrm{C}_p\sigma_i^2 \cdot |||\boldsymbol{Z}_{i-1}|||_{p,q}^2 \\
&\le |||\boldsymbol{Z}_{i-1} - \mathbb{E}\,\boldsymbol{Z}_{i-1}|||_{p,q}^2 + \mathrm{C}_p\sigma_i^2 \exp\left(\sum\nolimits_{k=1}^{i-1} \mathrm{C}_p\sigma_i^2\right) \cdot \|\boldsymbol{Z}_0\|_p^2 .
\end{aligned}$$

The last inequality is our growth bound (5.7). This recurrence relation delivers

$$\begin{aligned}
|||\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n|||_{p,q}^2 &\le |||\boldsymbol{Z}_0 - \mathbb{E}\,\boldsymbol{Z}_0|||_{p,q}^2 + \left[\sum\nolimits_{i=1}^{n} \mathrm{C}_p\sigma_i^2 \exp\left(\sum\nolimits_{k=1}^{i-1} \mathrm{C}_p\sigma_k^2\right)\right] \cdot \|\boldsymbol{Z}_0\|_p^2 \\
&= \left[\sum\nolimits_{i=1}^{n} \mathrm{C}_p\sigma_i^2 \exp\left(\sum\nolimits_{k=1}^{i-1} \mathrm{C}_p\sigma_k^2\right)\right] \cdot \|\boldsymbol{Z}_0\|_p^2 \\
&\le \left[\exp\left(\sum\nolimits_{i=1}^{n} \mathrm{C}_p\sigma_i^2\right) - 1\right] \cdot \|\boldsymbol{Z}_0\|_p^2 .
\end{aligned}$$

The equality holds because $\boldsymbol{Z}_0$ is not random. The last relation is a numerical inequality, whose proof appears in Lemma A.2. □

Observe that the difference between the bounds (5.5) and (5.6) is only visible when $\mathrm{C}_p v$ is small, in which case

$$\mathrm{e}^{\mathrm{C}_p v/2} \approx 1 \quad \text{and} \quad \left(\mathrm{e}^{\mathrm{C}_p v} - 1\right)^{1/2} \approx \sqrt{\mathrm{C}_p v}. \tag{5.8}$$

This is the setting where the concentration result may be nontrivial.

The next two remarks contain some minor extensions of Theorem 5.1. Similar extensions are possible at other points in this paper. For the most part, we omit these developments.

Remark 5.2 (Growth from Concentration)**.** In some instances, we can improve over the growth bound (5.5) by applying the triangle inequality to the decomposition $\boldsymbol{Z}_n = (\mathbb{E}\,\boldsymbol{Z}_n) + (\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n)$ and invoking the concentration bound (5.6):

$$|||\boldsymbol{Z}_n|||_{p,q} \le \|\mathbb{E}\,\boldsymbol{Z}_n\|_p + \left(\mathrm{e}^{\mathrm{C}_p v} - 1\right)^{1/2} \|\boldsymbol{Z}_0\|_p \cdot M.$$

Similarly, we can apply Proposition 4.3 together with (5.6) to obtain

$$|||\boldsymbol{Z}_n|||_{p,q}^2 \le \|\mathbb{E}\,\boldsymbol{Z}_n\|_p^2 + \mathrm{C}_p\left(\mathrm{e}^{\mathrm{C}_p v} - 1\right) \|\boldsymbol{Z}_0\|_p^2 \cdot M^2.$$

Neither of these bounds represents a strict improvement over the other or over the growth bound (5.5).

Remark 5.3 (Uniform Bounds on Factors)**.** Potentially stronger estimates are possible if the factors are bounded in norm. Fix parameters $2 \le q \le p$. Suppose that $\|\boldsymbol{Y}_i\| \le b_i$ almost surely and $|||\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i|||_{p,q} \le \sigma_i b_i$ for each index i. Define $B = \prod_{i=1}^n b_i$ and $v = \sum_{i=1}^n \sigma_i^2$. Then

$$|||\boldsymbol{Z}_n|||_{p,q} \le \|\boldsymbol{Z}_0\|_p \cdot B; \tag{5.9}$$

$$|||\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n|||_{p,q} \le \sqrt{\mathrm{C}_p v}\ \|\boldsymbol{Z}_0\|_p \cdot B. \tag{5.10}$$

Compare these results with (5.5), (5.6), and (5.8). As for the proof, the growth bound (5.9) is an immediate consequence of the definition $\boldsymbol{Z}_n = \boldsymbol{Y}_n \cdots \boldsymbol{Y}_1\boldsymbol{Z}_0$. The concentration result (5.10) follows if we repeat the proof of (5.6), using the growth bound (5.9) in place of (5.5).

5.3. **Expectation Bounds for the Spectral Norm.** In many cases, we just need to know the expected value of the product $\|\boldsymbol{Z}_n\|$ or the expected value of the fluctuation $\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\|$. We can obtain bounds for these quantities as an easy consequence of Theorem 5.1.

Corollary 5.4 (Expectation Bounds). *Consider an independent sequence* $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n\} \subset \mathbb{M}_d$ *of random matrices, and form the product* $\boldsymbol{Z}_n = \boldsymbol{Y}_n \cdots \boldsymbol{Y}_1$. *Assume that*

$$\|\mathbb{E}\,\boldsymbol{Y}_i\| \le m_i \quad \text{and} \quad \left(\mathbb{E}\,\|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i\|^2\right)^{1/2} \le \sigma_i m_i \quad \text{for } i = 1, \dots, n.$$

Let $M = \prod_{i=1}^n m_i$ *and* $v = \sum_{i=1}^n \sigma_i^2$. *Then*

$$\mathbb{E}\,\|\boldsymbol{Z}_n\| \le \exp\left(\sqrt{2v\,(2v \vee \log d)}\right) \cdot M. \tag{5.11}$$

Provided that $v\,(1 + 2\log d) \le 1$, *then also*

$$\mathbb{E}\,\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \le \sqrt{\mathrm{e}^2 v\,(1 + 2\log d)} \cdot M. \tag{5.12}$$

Proof. To apply Theorem 5.1, we set $\boldsymbol{Z}_0 = \mathbf{I}$ and choose the power $q = 2$.

To obtain the growth bound (5.11), consider the Schatten norm of order $p = \sqrt{2(2v \vee \log d)/v}$. Note that $p \ge 2$ and that $\|\boldsymbol{Z}_0\|_p \le d^{1/p} \le \mathrm{e}^{pv/2}$. Invoke Theorem 5.1, relation (5.5), to see that

$$\mathbb{E}\,\|\boldsymbol{Z}_n\| \le \|\!|\boldsymbol{Z}_n|\!\|_{p,2} \le \mathrm{e}^{\mathrm{C}_p v/2}\,\|\boldsymbol{Z}_0\|_p \cdot M \le \mathrm{e}^{pv/2} \cdot \mathrm{e}^{pv/2} \cdot M = \mathrm{e}^{pv} \cdot M.$$

We used the fact that $\mathrm{C}_p = p - 1 < p$. This is the stated result.

To obtain the concentration bound (5.12), consider the Schatten norm $p = 2(1 + \log d)$. Note that $p \ge 2$ and that $\|\boldsymbol{Z}_0\|_p \le d^{1/p} \le \sqrt{\mathrm{e}}$. Now, we use Theorem 5.1, relation (5.6), in a similar fashion. Assuming that $\mathrm{C}_p v \le 1$,

$$\mathbb{E}\,\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \le \|\!|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n|\!\|_{p,2} \le \left(\mathrm{e}^{\mathrm{C}_p v} - 1\right)^{1/2} \|\boldsymbol{Z}_0\|_p \cdot M \le \mathrm{e}\sqrt{\mathrm{C}_p v} \cdot M.$$

The last bound is the numerical inequality $\mathrm{e}^a - 1 \le \mathrm{e}a$, valid when $a \in [0, 1]$. Finally, note that $\mathrm{C}_p = p - 1 = 1 + 2\log d$. ∎

The inequality (5.11) shows its power when each σ_i is small. Assume that each $m_i = 1$ and $\sigma_i \le b/n$ for a constant b. Then it is not hard to check that

$$\|\mathbb{E}\,\boldsymbol{Z}_n\| \le 1 \quad \text{and} \quad \|\boldsymbol{Z}_n\| \le (1 + (b/n))^n \le \mathrm{e}^b.$$

If $L\sqrt{(2\log d)/n}$ is close to zero, then (5.11) implies

$$\mathbb{E}\,\|\boldsymbol{Z}_n\| \le \mathrm{e}^{b\sqrt{(2\log d)/n}} \approx 1.$$

That is, $\mathbb{E}\,\|\boldsymbol{Z}_n\|$ is much closer to $\|\mathbb{E}\,\boldsymbol{Z}_n\|$ than to the worst-case value e^b.

Remark 5.5 (Uniform Bounds on Factors). Fix $p \ge 2$. Assume that $\|\boldsymbol{Y}_i\| \le b_i$ almost surely and $\|\!|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i|\!\|_{p,2} \le \sigma_i b_i$ for each i. Let $v = \sum_{i=1}^n \sigma_i^2$ and $B = \prod_{i=1}^n b_i$. Then Remark 5.3 implies that

$$\mathbb{E}\,\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \le \sqrt{\mathrm{e}v\,(1 + 2\log d)}B.$$

This improves the constant in (5.12) by a factor of $\sqrt{\mathrm{e}}$, and it removes the condition that $v\,(1+2\log d) \le 1$.

5.4. **Tail Bounds for the Spectral Norm.** The moment bounds in Theorem 5.1 can also be upgraded to obtain tail bounds for $\|\boldsymbol{Z}_n\|$ and $\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\|$.

Corollary 5.6 (Tail Bounds). *Consider an independent sequence* $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n\} \subset \mathbb{M}_d$ *of random matrices, and form the product* $\boldsymbol{Z}_n = \boldsymbol{Y}_n \cdots \boldsymbol{Y}_1$. *Assume that*

$$\|\mathbb{E}\,\boldsymbol{Y}_i\| \le m_i \quad \text{and} \quad \|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i\| \le \sigma_i m_i \quad \text{almost surely for } i = 1, \dots, n.$$

Let $M = \prod_{i=1}^n m_i$ *and* $v = \sum_{i=1}^n \sigma_i^2$. *Then*

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n\| \ge tM\right\} \le d \cdot \exp\left(\frac{-\log^2 t}{2v}\right) \quad \text{when } \log t \ge 2v. \tag{5.13}$$

Furthermore,

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \geq tM\right\} \leq (d \vee \mathrm{e}) \cdot \exp\left(\frac{-t^2}{2\mathrm{e}^2 v}\right) \quad \textit{when } t \leq \mathrm{e}. \tag{5.14}$$

Proof. We begin with the proof of (5.13). By homogeneity, we may assume that $m_i = 1$ for each i, so also $M = 1$. Apply Markov's inequality and (4.2) to obtain

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n\| \geq t\right\} \leq \inf_{p \geq 2} t^{-p} \cdot \mathbb{E}\,\|\boldsymbol{Z}_n\|^p \leq \inf_{p \geq 2} t^{-p} \cdot |||\boldsymbol{Z}_n|||_{p,p}^p.$$

To bound the $L_p(S_p)$ norm, we will use Theorem 5.1 with $\boldsymbol{Z}_0 = \mathbf{I}$ and with $q = p$. Relation (5.5) gives

$$t^{-p} \cdot |||\boldsymbol{Z}_n|||_{p,p}^p \leq t^{-p} \cdot \mathrm{e}^{p\mathrm{C}_p v/2}\,\|\boldsymbol{Z}_0\|_p^p = d \cdot \left(t^{-2}\mathrm{e}^{\mathrm{C}_p v}\right)^{p/2}.$$

We have used the fact that $\|\boldsymbol{Z}_0\|_p^p = \|\mathbf{I}\|_p^p = d$. Under the assumption that $\log t \geq 2v$, we may select $p = (\log t)/v \geq 2$. This choice yields

$$d \cdot \left(t^{-2}\mathrm{e}^{pv}\right)^{p/2} = d \cdot \exp\left(\frac{-\log^2 t}{2v}\right).$$

Sequence the last three displays to arrive at the bound (5.13).

We establish (5.14) in an analogous fashion. The same argument, using relation (5.6), implies that

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \geq t\right\} \leq \inf_{p \geq 2} d \cdot \left[t^{-2}\left(\mathrm{e}^{\mathrm{C}_p v} - 1\right)\right]^{p/2}.$$

Supposing that $t^2/(\mathrm{e}^2 v) < 2$, the bound (5.14) holds trivially because $\mathrm{e} \cdot \exp(-t^2/(2\mathrm{e}^2 v)) \geq 1$. Otherwise, we may select the parameter $p = t^2/(\mathrm{e}^2 v) \geq 2$. Under the assumption that $t \leq \mathrm{e}$, $\mathrm{C}_p v \leq pv \leq (t/\mathrm{e})^2 \leq 1$, so that $\mathrm{e}^{\mathrm{C}_p v} - 1 \leq \mathrm{e}\mathrm{C}_p v \leq t^2/\mathrm{e}$. Therefore,

$$d \cdot \left[t^{-2}\left(\mathrm{e}^{\mathrm{C}_p v} - 1\right)\right]^{p/2} \leq d \cdot \mathrm{e}^{-p/2} = d \cdot \exp\left(\frac{-t^2}{2\mathrm{e}^2 v}\right).$$

The last two displays imply (5.14). □

Remark 5.7 (Uniform Bounds on Factors)**.** In the setting of Remark 5.5, we have an unconditional variant of the concentration bound (5.14):

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \geq t \cdot B\right\} \leq (d \vee \mathrm{e}) \cdot \exp\left(\frac{-t^2}{2\mathrm{e}v}\right) \quad \text{for all } t > 0.$$

6. Application: Random Perturbations of the Identity

This section treats the fundamental case where the factors $\boldsymbol{Y}_i$ in the product are independent, random perturbations of the identity. That is, $\boldsymbol{Y}_i = \mathbf{I} + \boldsymbol{X}_i$ where $\{\boldsymbol{X}_i\} \subset \mathbb{M}_d$ is an independent family. We will develop specialized theory for this class of problems, and we will use these results to compare our work with several recent papers.

6.1. Iterative Algorithms.

To motivate this development, observe that random perturbations of the identity arise from the analysis of the iterative scheme

$$\boldsymbol{u}^{(i+1)} = \boldsymbol{u}^{(i)} + \boldsymbol{X}_i\boldsymbol{u}^{(i)} \quad \text{for } i = 1, 2, 3 \ldots. \tag{6.1}$$

where $\boldsymbol{X}_i\boldsymbol{u}^{(i)}$ is a linear update to the current iterate $\boldsymbol{u}^{(i)}$. In this application, the norm of each $\boldsymbol{X}_i$ is proportional to the step size of the scheme, so it is typically small and it is controlled by the user. For example, the updates in Oja's algorithm [28] take the form (6.1).

For now, we do not permit the random matrix $\boldsymbol{X}_i$ to depend on the sequence $\{\boldsymbol{u}^{(i)}\}$ of iterates. Later, in Section 7.3, we describe an extension of our approach to the setting where $\{\boldsymbol{X}_i\}$ is an adapted sequence. This variant allows for the study of a wider class of iterative algorithms.

6.2. **Bounds for the Product.** First, we develop bounds for the growth and concentration of a product of perturbations of the identity. In Section 6.4, we develop results for the inverse of the product.

Corollary 6.1 (Perturbations of the Identity)**.** *Consider an independent family $\{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\} \subset \mathbb{M}_d$ of random matrices, and form the product $\boldsymbol{Z}_n = (\mathbf{I} + \boldsymbol{X}_n) \cdots (\mathbf{I} + \boldsymbol{X}_1)$. Assume that*

$$\|\mathbb{E}\,\boldsymbol{X}_i\| \le \xi_i \quad \text{and} \quad \|\boldsymbol{X}_i - \mathbb{E}\,\boldsymbol{X}_i\| \le \sigma_i \quad \text{almost surely for } i = 1, \dots, n.$$

Define $\xi = \sum_{i=1}^n \xi_i$ and $v = \sum_{i=1}^n \sigma_i^2$. Then

$$\mathbb{E}\,\|\boldsymbol{Z}_n\| \le \exp\left(\xi + \sqrt{2v \log d}\right) \qquad \text{when } 2v \le \log d;$$

$$\mathbb{E}\,\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \le \mathrm{e}^{\xi+1}\sqrt{v\,(1 + 2\log d)} \qquad \text{when } v\,(1 + 2\log d) \le 1.$$

Moreover,

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n\| \ge t\mathrm{e}^{\xi}\right\} \le d \cdot \exp\left(\frac{-\log^2 t}{2v}\right) \qquad \text{when } \log t \ge 2v;$$

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \ge t\mathrm{e}^{\xi}\right\} \le (d \vee \mathrm{e}) \cdot \exp\left(\frac{-t^2}{2\mathrm{e}^2 v}\right) \qquad \text{when } t \le \mathrm{e}.$$

Proof. Let $\boldsymbol{Y}_i = \mathbf{I} + \boldsymbol{X}_i$ for each index i. Then

$$\|\mathbb{E}\,\boldsymbol{Y}_i\| \le 1 + \|\mathbb{E}\,\boldsymbol{X}_i\| \le \mathrm{e}^{\xi_i} =: m_i.$$

Furthermore, since $m_i \ge 1$,

$$\|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i\| = \|\boldsymbol{X}_i - \mathbb{E}\,\boldsymbol{X}_i\| \le \sigma_i \le \sigma_i m_i.$$

The results follow instantly from Corollary 5.4 and Corollary 5.6. □

6.3. **Comparison with Prior Work.** To clarify the meaning of Corollary 6.1, let us elaborate what it predicts when

$$\|\mathbb{E}\,\boldsymbol{X}_i\| \le T/n \quad \text{and} \quad \|\boldsymbol{X}_i - \mathbb{E}\,\boldsymbol{X}_i\| \le L/n \quad \text{for constants } T, L.$$

This situation can arise if we perform n iterations of the iterative scheme (6.1) with a uniform step size of $1/n$. In this setting, Corollary 6.1 implies that

$$\mathbb{E}\,\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \le \sqrt{\frac{1 + 2\log d}{n}}\,L\mathrm{e}^{1+T} \quad \text{when } L^2(1 + 2\log d) \le n. \tag{6.2}$$

For $\delta \in [0, 1]$, with probability at least $1 - \delta$,

$$\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \le \sqrt{\frac{2 + 2\log(d/\delta)}{n}}\,L\mathrm{e}^{1+T} \quad \text{when } L^2(2 + 2\log(d/\delta)) \le n. \tag{6.3}$$

Furthermore, if we assume that $\|\boldsymbol{X}_i\| \le T/n$ almost surely for each i, then Remark 5.5 implies that (6.2) and (6.3) hold without restriction.

The paper [19] of Henriksen and Ward only contemplates the situation described in the last paragraph. It obtains a concentration bound of the form

$$\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \le \frac{L\mathrm{e}^L}{\sqrt{n}} \cdot \mathrm{polylog}(n, d, 1/\delta) \quad \text{with probability at least } 1 - \delta.$$

The salient improvement in (6.3) stems from the reduction of the factor e^L to e^T. This difference is most pronounced when $\mathbb{E}\,\boldsymbol{X}_i = 0$ for each i, in which case the bound (6.3) removes the exponential

factor entirely. Even under the assumption that $\boldsymbol{X}_i \succcurlyeq \mathbf{0}$ for all each i, it can happen that $L \geq dT$, so this refinement can make a big difference.

Last, we mention one instance that has special importance. Let $\boldsymbol{A} \in \mathbb{M}_d$ be a fixed matrix. Consider a triangular array $\{\boldsymbol{X}_i^{(n)} : i \leq n \text{ and } n \in \mathbb{N}\} \subset \mathbb{M}_d$ of independent random matrices. For each index n, assume that

$$\mathbb{E}\,\boldsymbol{X}_i^{(n)} = \boldsymbol{A}/n \quad \text{and} \quad \|\boldsymbol{X}_i^{(n)} - \mathbb{E}\,\boldsymbol{X}_i^{(n)}\| \leq L/n \quad \text{for } i = 1, \dots, n.$$

Define the product

$$\boldsymbol{Z}^{(n)} = \left(\mathbf{I} + \boldsymbol{X}_n^{(n)}\right) \cdots \left(\mathbf{I} + \boldsymbol{X}_1^{(n)}\right).$$

By functional calculus,

$$\mathbb{E}\,\boldsymbol{Z}^{(n)} = (\mathbf{I} + \boldsymbol{A}/n)^n \to \mathrm{e}^{\boldsymbol{A}} \quad \text{as } n \to \infty.$$

The bound (6.3), combined with the first Borel–Cantelli Lemma, guarantees that

$$\boldsymbol{Z}^{(n)} \to \mathrm{e}^{\boldsymbol{A}} \quad \text{as } n \to \infty, \text{ almost surely.}$$

This result is a special case of the limit theorem of Emme and Hubert [13, Thm. 1.1]. They do not require independence, but they only achieve an asymptotic result. Our analysis gives a rate of convergence that matches the corresponding bound (2.2) for scalar random variables.

6.4. **Bounds for the Inverse of a Product.** In some applications, it is valuable to have a lower bound for the minimum singular value of a random product. Equivalently, we can seek an upper bound for the spectral norm of the inverse of the product. This section describes a situation where clean results are possible.

Consider the case where the factors $\boldsymbol{Y}_i$ are perturbations of the identity: $\boldsymbol{Y}_i = \mathbf{I} + \boldsymbol{X}_i$, where $\boldsymbol{X}_i$ is small enough to ensure that $\boldsymbol{Y}_i$ is invertible with probability 1. In this setting, we can easily study the inverse of the product using Corollary 6.1.

Corollary 6.2 (Perturbations of the Identity: Inverses)**.** *Frame the same hypotheses as in Corollary 6.1. Assume that $\xi_i + \sigma_i < 1$ for each index i, and define*

$$\bar{\xi} = \sum\nolimits_{i=1}^n \left[\xi_i + \frac{(\xi_i + \sigma_i)^2}{1 - (\xi_i + \sigma_i)}\right] \quad \text{and} \quad \bar{v} = \sum\nolimits_{i=1}^n \left[\sigma_i + \frac{2(\xi_i + \sigma_i)^2}{1 - (\xi_i + \sigma_i)}\right]^2.$$

Then

$$\mathbb{E}\left\|\boldsymbol{Z}_n^{-1}\right\| \leq \exp\left(\bar{\xi} + \sqrt{2\bar{v}\log d}\right) \qquad \text{when } 2\bar{v} \leq \log d;$$

$$\mathbb{E}\left\|\boldsymbol{Z}_n^{-1} - \mathbb{E}\,\boldsymbol{Z}_n^{-1}\right\| \leq \mathrm{e}^{\bar{\xi}}\sqrt{\mathrm{e}^2\bar{v}\,(1 + 2\log d)} \qquad \text{when } \bar{v}\,(1 + 2\log d) \leq 1.$$

Proof. With the same notation as in Corollary 6.1, observe that $\boldsymbol{Z}_n^{-1} = (\mathbf{I} + \boldsymbol{X}_1)^{-1} \cdots (\mathbf{I} + \boldsymbol{X}_n)^{-1}$. This is an independent product that can be bounded by applying the corollary. To do so, we simply need to express $(\mathbf{I} + \boldsymbol{X}_i)^{-1} = \mathbf{I} + \bar{\boldsymbol{X}}_i$ for suitable random matrices $\bar{\boldsymbol{X}}_i$. The perturbation terms $\bar{\boldsymbol{X}}_i$ are obtained from the calculation

$$(\mathbf{I} + \boldsymbol{X}_i)^{-1} = \mathbf{I} + \sum\nolimits_{k=1}^{\infty}(-1)^k\boldsymbol{X}_i^k = \mathbf{I} - \boldsymbol{X}_i + \boldsymbol{X}_i^2(\mathbf{I} + \boldsymbol{X}_i)^{-1} =: \mathbf{I} + \bar{\boldsymbol{X}}_i.$$

It remains to develop estimates for the size of the perturbation.

The uniform bound $\|\boldsymbol{X}_i\| \leq \|\mathbb{E}\,\boldsymbol{X}_i\| + \|\boldsymbol{X}_i - \mathbb{E}\,\boldsymbol{X}_i\| \leq \xi_i + \sigma_i < 1$ implies that

$$\left\|(\mathbf{I} + \boldsymbol{X}_i)^{-1}\right\| \leq \left(1 - \|\boldsymbol{X}_i\|\right)^{-1} \leq \frac{1}{1 - (\xi_i + \sigma_i)}.$$

Therefore, the norm of the expected perturbation satisfies

$$\left\|\mathbb{E}\,\bar{\boldsymbol{X}}_i\right\| \leq \|\mathbb{E}\,\boldsymbol{X}_i\| + \left\|\mathbb{E}\left[\boldsymbol{X}_i^2(\mathbf{I} + \boldsymbol{X}_i)^{-1}\right]\right\| \leq \xi_i + \frac{(\xi_i + \sigma_i)^2}{1 - (\xi_i + \sigma_i)} =: \bar{\xi}_i.$$

The fluctuations of the perturbation satisfy

$$\left\|\bar{\boldsymbol{X}}_i - \mathbb{E}\,\bar{\boldsymbol{X}}_i\right\| \le \|\boldsymbol{X}_i - \mathbb{E}\,\boldsymbol{X}_i\| + 2\left\|\boldsymbol{X}_i^2(\mathbf{I} + \boldsymbol{X}_i)^{-1}\right\| \le \sigma_i + \frac{2(\xi_i + \sigma_i)^2}{1 - (\xi_i + \sigma_i)} =: \bar{\sigma}_i.$$

The results follow when we apply Corollary 6.1 with the random matrices $\bar{\boldsymbol{X}}_i$ in place of the $\boldsymbol{X}_i$. □

7. Improvements and Extensions

The argument underlying Theorem 5.1 has several natural extensions. First, we develop sharper results for products of random contractions. In Section 7.2, we derive better estimates for a matrix product where the initial term is rectangular. In Section 7.3, we document the changes that are necessary in case the factors in the product are not independent but form an adapted sequence. Last, In Section 7.4, we explain how to develop a bound on the spectral radius of a product.

7.1. A Product of Contractions.

Most of our results are designed for products of general random matrices. In some circumstances, the factors in the product are *contractions*, matrices whose singular values are bounded by one. For example, the randomized Kaczmarz algorithm [36] can be expressed as the repeated application of random contractions. Other randomized linear fixed-point iterations take a similar form. This section derives sharper estimates for this important setting.

Theorem 7.1 (Product of Contractions). *Consider an independent family $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n\} \subset \mathbb{M}_d$ of random contractions; that is, $\|\boldsymbol{Y}_i\| \le 1$. Form the random product $\boldsymbol{Z}_n = \boldsymbol{Y}_n \cdots \boldsymbol{Y}_1$. Assume that*

$$\left\|\mathbb{E}\,|\boldsymbol{Y}_i|^2\right\| \le m_i^2 \le 1 \quad \text{and} \quad \|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i\| \le \sigma_i m_i \quad \text{almost surely for } i = 1, \dots, n.$$

Define $M := \prod_{i=1}^n m_i$ and $v := \sum_{i=1}^n \sigma_i^2$. Then

$$\mathbb{E}\,\|\boldsymbol{Z}_n\| \le 1 \wedge (\sqrt{d} \cdot M); \tag{7.1}$$

$$\mathbb{E}\,\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \le \sqrt{dv} \cdot M. \tag{7.2}$$

Furthermore, we have the tail bound

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \ge t\right\} \le dM^2 \cdot \mathrm{e}^{-t^2/(2ev)} \quad \text{when } t^2 \ge 2ev. \tag{7.3}$$

To prove this result, we require a lemma that isolates the influence of each factor in the product. This step exploits the uniform bound on the singular values in an essential way.

Lemma 7.2 (Random Contractions). *Let $\boldsymbol{Y} \in \mathbb{M}_d$ be a random contraction, and let $\boldsymbol{Z} \in \mathbb{M}_d$ be a random matrix that is independent from $\boldsymbol{Y}$. For $2 \le q \le p$,*

$$|||\boldsymbol{YZ}|||_{p,q} \le \left\|\mathbb{E}\,|\boldsymbol{Y}|^2\right\|^{1/p} \cdot |||\boldsymbol{Z}|||_{p,q}.$$

Proof. Write out the $L_q(S_p)$ norm, and introduce matrix absolute values:

$$|||\boldsymbol{YZ}|||_{p,q}^q = \mathbb{E}\,\|\boldsymbol{YZ}\|_p^q = \mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{Z}^*\boldsymbol{Y}^*\boldsymbol{YZ}\right)^{p/2}\right]^{q/p} = \mathbb{E}\left[\operatorname{tr}\left(|\boldsymbol{Z}^*| \cdot |\boldsymbol{Y}|^2 \cdot |\boldsymbol{Z}^*|\right)^{p/2}\right]^{q/p}.$$

The last relation can be verified using polar factorizations. Apply the Araki–Lieb–Thirring inequality [9, Thm. IX.2.20] to distribute the power onto the factors in the trace. We obtain

$$\begin{aligned}
|||\boldsymbol{YZ}|||_{p,q}^q &\le \mathbb{E}\left[\operatorname{tr}\left(|\boldsymbol{Z}^*|^{p/2} \cdot |\boldsymbol{Y}|^p \cdot |\boldsymbol{Z}^*|^{p/2}\right)\right]^{q/p} \\
&\le \mathbb{E}_{\boldsymbol{Z}}\,\mathbb{E}_{\boldsymbol{Y}}\left[\operatorname{tr}\left(|\boldsymbol{Z}^*|^{p/2} \cdot |\boldsymbol{Y}|^2 \cdot |\boldsymbol{Z}^*|^{p/2}\right)\right]^{q/p} \\
&\le \mathbb{E}_{\boldsymbol{Z}}\left[\operatorname{tr}\left(|\boldsymbol{Z}^*|^{p/2} \cdot \left(\mathbb{E}_{\boldsymbol{Y}}\,|\boldsymbol{Y}|^2\right) \cdot |\boldsymbol{Z}^*|^{p/2}\right)\right]^{q/p}.
\end{aligned}$$

The second inequality holds because a contraction satisfies $|\boldsymbol{Y}|^p \preccurlyeq |\boldsymbol{Y}|^2$ for each $p \geq 2$. The third inequality is Jensen's, justified because $q/p \leq 1$. Bounding the matrix in the center by its norm,

$$|||\boldsymbol{Y}\boldsymbol{Z}|||_{p,q}^q \leq \left\|\mathbb{E}\,|\boldsymbol{Y}|^2\right\|^{q/p} \cdot \mathbb{E}\left[\operatorname{tr}|\boldsymbol{Z}^*|^p\right]^{q/p} = \left\|\mathbb{E}\,|\boldsymbol{Y}|^2\right\|^{q/p} \cdot |||\boldsymbol{Z}|||_{p,q}^q.$$

This completes the analysis. □

With this result at hand, Theorem 7.1 follows from familiar arguments.

Proof of Theorem 7.1. Define $\boldsymbol{Z}_0 = \mathbf{I}$ and $\boldsymbol{Z}_i = \boldsymbol{Y}_i\boldsymbol{Z}_{i-1}$ for each index $i = 1, \dots, n$. We begin with the proof of (7.1). Since each factor is a contraction, it is clear that

$$\mathbb{E}\,\|\boldsymbol{Z}_n\| \leq \mathbb{E}\prod\nolimits_{k=1}^n \|\boldsymbol{Y}_k\| \leq 1.$$

To obtain a less trivial bound on the expectation, we apply Lemma 7.2 repeatedly. For $p \geq 2$,

$$\mathbb{E}\,\|\boldsymbol{Z}_i\| \leq |||\boldsymbol{Z}_i|||_{p,p} \leq \prod\nolimits_{k=1}^i \left\|\mathbb{E}\,|\boldsymbol{Y}_k|^2\right\|^{1/p} \cdot |||\mathbf{I}|||_{p,p} \leq d^{1/p}\prod\nolimits_{k=1}^i m_k^{2/p}. \tag{7.4}$$

The statement (7.1) combines these two observations when we set $i = n$ and $p = 2$.

Let us continue with the proof of (7.2), which is analogous to the argument in Theorem 5.1(5.6). First, by expanding the inequality $\mathbb{E}\,|\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i|^2 \succcurlyeq \mathbf{0}$, we see that $\mathbf{0} \preccurlyeq |\mathbb{E}\,\boldsymbol{Y}_i|^2 \preccurlyeq \mathbb{E}\,|\boldsymbol{Y}_i|^2$. As a consequence,

$$\|\mathbb{E}\,\boldsymbol{Y}_i\|^2 \leq \left\|\mathbb{E}\,|\boldsymbol{Y}_i|^2\right\| \leq m_i^2.$$

For $p \geq 2$, calculate that

$$\begin{aligned}
|||\boldsymbol{Z}_i - \mathbb{E}\,\boldsymbol{Z}_i|||_{p,p}^2 &\leq |||(\mathbb{E}\,\boldsymbol{Y}_i)(\boldsymbol{Z}_{i-1} - \mathbb{E}\,\boldsymbol{Z}_{i-1})|||_{p,p}^2 + \mathrm{C}_p \cdot |||(\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{Z}_{i-1}|||_{p,p}^2 \\
&\leq m_i^2 \cdot |||\boldsymbol{Z}_{i-1} - \mathbb{E}\,\boldsymbol{Z}_{i-1}|||_{p,p}^2 + \mathrm{C}_p\sigma_i^2 m_i^2 \cdot |||\boldsymbol{Z}_{i-1}|||_{p,p}^2 \\
&\leq m_i^{4/p} \cdot |||\boldsymbol{Z}_{i-1} - \mathbb{E}\,\boldsymbol{Z}_{i-1}|||_{p,p}^2 + \mathrm{C}_p\sigma_i^2 \cdot d^{2/p}\prod\nolimits_{k=1}^i m_k^{4/p}.
\end{aligned}$$

The second inequality is Lemma 7.2, and the third inequality requires (7.4). We have also used the fact that $m_i^2 \leq m_i^{4/p}$ because $m_i \leq 1$. Unrolling the recursion,

$$|||\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n|||_{p,p}^2 \leq \mathrm{C}_p d^{2/p}\left(\prod\nolimits_{i=1}^n m_i^{4/p}\right)\left(\sum\nolimits_{i=1}^n \sigma_i^2\right) = \mathrm{C}_p d^{2/p} M^{4/p} v. \tag{7.5}$$

For $p = 2$, this result implies the advertised bound (7.2).

Finally, the tail inequality (7.3) follows from the estimate

$$\mathbb{P}\left\{\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \geq t\right\} \leq \min_{p\geq 2}\, t^{-p} \cdot |||\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n|||_{p,p}^p \leq (dM^2) \cdot \min_{p\geq 2}\left(\frac{pv}{t^2}\right)^{p/2}.$$

The last inequality follows from (7.5) and $\mathrm{C}_p < p$. Bound the minimum with the power $p = t^2/(\mathrm{e}v) \geq 2$ to complete the argument. □

7.2. **Low-Rank Products.** So far, we have focused on the setting where the initial matrix $\boldsymbol{Z}_0 = \mathbf{I}$. In many applications, we are interested in the action of the random product $\boldsymbol{Y}_n \cdots \boldsymbol{Y}_1 \in \mathbb{M}_d$ on a specific matrix $\boldsymbol{Z}_0 \in \mathbb{C}^{d\times r}$ with relatively few columns. In this case, the terms that the control the behavior of the product may be significantly smaller. Here is an example of the kinds of results one can achieve.

Theorem 7.3 (Growth and Concentration of Low-Rank Products)**.** *Consider a fixed matrix $\boldsymbol{Z}_0 \in \mathbb{C}^{d\times r}$ and an independent sequence $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n\} \subset \mathbb{M}_d$ of random matrices. Form the product $\boldsymbol{Z}_n = \boldsymbol{Y}_n \cdots \boldsymbol{Y}_1\boldsymbol{Z}_0$. Assume that*

$$\|\mathbb{E}\,\boldsymbol{Y}_i\| \leq m_i \quad \text{and} \quad \sup_{\boldsymbol{P}\in\mathscr{P}_r}\left(\mathbb{E}\,\|(\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{P}\|^2\right)^{1/2} \leq \sigma_i m_i \quad \text{for } i = 1, \dots, n,$$

where $\mathcal{P}_r \subset \mathbb{M}_d$ is the set of rank-r orthogonal projectors. Define $M = \prod_{i=1}^n m_i$ and $v = \sum_{i=1}^n \sigma_i^2$. For each $p \geq 2$,

$$\mathbb{E}\,\|\boldsymbol{Z}_n\| \leq \mathrm{e}^{\mathrm{C}_p v/2} \cdot \|\boldsymbol{Z}_0\|_p \cdot M.$$
$$\mathbb{E}\,\|\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{Z}_n\| \leq \left(\mathrm{e}^{\mathrm{C}_p v} - 1\right)^{1/2} \cdot \|\boldsymbol{Z}_0\|_p \cdot M.$$

Proof. Define $\boldsymbol{Z}_i = \boldsymbol{Y}_i \boldsymbol{Z}_{i-1}$ for each index i. Since $\boldsymbol{Z}_0 \in \mathbb{C}^{d\times r}$, the rank of each matrix $\boldsymbol{Z}_i$ is at most r. Thus, we can write $\boldsymbol{Z}_i = \boldsymbol{P}_i \boldsymbol{Z}_i$, where $\boldsymbol{P}_i$ is a rank-r orthogonal projector that only depends on $\boldsymbol{Y}_i, \dots, \boldsymbol{Y}_1$ and $\boldsymbol{Z}_0$. As a consequence,

$$\begin{aligned}
\|\!|(\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{Z}_{i-1}|\!\|_{p,2} &= \|\!|(\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{P}_{i-1}\boldsymbol{Z}_{i-1}|\!\|_{p,2} \\
&\leq \left(\mathbb{E}\left[\,\|(\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{P}_{i-1}\|^2 \cdot \|\boldsymbol{Z}_{i-1}\|_p^2\,\right]\right)^{1/2} \\
&\leq \sup_{\boldsymbol{P}\in\mathcal{P}_r} \left(\mathbb{E}\,\|(\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i)\boldsymbol{P}\|^2\right)^{1/2} \cdot \left(\mathbb{E}\,\|\boldsymbol{Z}_{i-1}\|_p^2\right)^{1/2} \leq \sigma_i m_i \cdot \|\!|\boldsymbol{Z}_{i-1}|\!\|_{p,2}.
\end{aligned}$$

We have used the fact that $\boldsymbol{Y}_i$ is independent from $\boldsymbol{P}_{i-1}$ and from $\boldsymbol{Z}_{i-1}$ to pass to the last line.

The rest of the proof runs along the same lines as the argument in Theorem 5.1, using the last display in place of the bound (5.3). ∎

Let us offer a simple example to illustrate why Theorem 7.3 can produce better outcomes than Theorem 5.1. Consider a random matrix $\boldsymbol{X} \in \mathbb{M}_d$ with the distribution $\mathbb{P}\{\boldsymbol{X} = \mathbf{e}_j \mathbf{e}_j^*\} = d^{-1}$ for each $j = 1, \dots, d$. As usual, $\mathbf{e}_j \in \mathbb{C}^d$ is the jth standard basis vector. Construct the random matrix $\boldsymbol{Y} = \mathbf{I} + \varepsilon \boldsymbol{X}$, where ε is a Rademacher random variable that is independent from $\boldsymbol{X}$. Clearly, $\mathbb{E}\,\boldsymbol{Y} = \mathbf{I}$. For any rank-$r$ orthogonal projector $\boldsymbol{P}$,

$$\mathbb{E}\,\|(\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y})\boldsymbol{P}\|^2 = \mathbb{E}\,\|\boldsymbol{P}\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{P}\| = \frac{1}{d}\sum\nolimits_{i=1}^d \operatorname{tr}[\boldsymbol{P}\mathbf{e}_i\mathbf{e}_i^*\boldsymbol{P}] = \frac{1}{d}\operatorname{tr}\boldsymbol{P} = \frac{r}{d}.$$

Therefore,

$$\sup_{\boldsymbol{P}\in\mathcal{P}_r} \left(\mathbb{E}\,\|(\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y})\boldsymbol{P}\|^2\right)^{1/2} = \sqrt{r/d} \leq 1.$$

By contrast, $\mathbb{E}\,\|\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}\|^2 = \mathbb{E}\,\|\boldsymbol{X}\|^2 = 1$. When $r \ll d$, this bound offers a significant improvement. instead of the ambient dimension d.

7.3. **Adapted Sequences.** We can easily generalize our results on a product of independent random matrices to a product of adapted random matrices. This kind of extension is valuable for studying iterative algorithms where the choices made by the algorithm at a given step depend on the history of the iteration.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_n \subset \mathcal{F}$ be a filtration. For each index $i = 1, \dots, n$, we write $\mathbb{E}_i$ for the expectation conditioned on the σ-algebra $\mathcal{F}_i$. The operator $\mathbb{E}_0 := \mathbb{E}$ is the unconditional expectation.

We consider an adapted sequence $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n\} \subset \mathbb{M}_d$ of random matrices; that is, each $\boldsymbol{Y}_i$ is measurable with respect to $\mathcal{F}_i$. The next result provides information about the growth and concentration properties of the product $\boldsymbol{Z}_n = \boldsymbol{Y}_n \cdots \boldsymbol{Y}_1$. Note that the natural concentration result compares $\boldsymbol{Z}_n$ with a product of conditional expectations, rather than the expectation of the product.

Theorem 7.4 (Products of Adapted Random Matrices)**.** *Consider a fixed matrix $\boldsymbol{Z}_0 \in \mathbb{M}_d$ and an adapted sequence $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n\} \subset \mathbb{M}_d$ of random matrices. Form the products*

$$\boldsymbol{Z}_n = \boldsymbol{Y}_n \cdots \boldsymbol{Y}_1 \boldsymbol{Z}_0 \quad \text{and} \quad \boldsymbol{F}_n = (\mathbb{E}_{n-1}\boldsymbol{Y}_n) \cdots (\mathbb{E}_1 \boldsymbol{Y}_2)(\mathbb{E}_0 \boldsymbol{Y}_1)\boldsymbol{Z}_0.$$

Assume that

$$\|\mathbb{E}_{i-1}\boldsymbol{Y}_i\| \leq m_i \quad \text{and} \quad \|\boldsymbol{Y}_i - \mathbb{E}_{i-1}\boldsymbol{Y}_i\| \leq \sigma_i m_i \quad \text{almost surely for } i = 1, \dots, n.$$

Define $M = \prod_{i=1}^{n} m_i$ *and* $v = \sum_{i=1}^{n} \sigma_i^2$. *For* $2 \le q \le p$, *the random product* $\boldsymbol{Z}_n$ *satisfies the growth and concentration bounds*

$$\|\!|\!|\boldsymbol{Z}_n|\!|\!|_{p,q} \le \mathrm{e}^{\mathrm{C}_p v/2} \, \|\boldsymbol{Z}_0\|_p \cdot M; \tag{7.6}$$

$$\|\!|\!|\boldsymbol{Z}_n - \boldsymbol{F}_n|\!|\!|_{p,q} \le \left(\mathrm{e}^{\mathrm{C}_p v} - 1\right)^{1/2} \|\boldsymbol{Z}_0\|_p \cdot M. \tag{7.7}$$

Proof. Recursively construct the products

$$\boldsymbol{Z}_i = \boldsymbol{Y}_i \boldsymbol{Z}_{i-1} \quad \text{and} \quad \boldsymbol{F}_i = (\mathbb{E}_{i-1} \boldsymbol{Y}_i)\boldsymbol{Z}_{i-1} \quad \text{for } i = 1, \dots, n.$$

To bound the growth of $\boldsymbol{Z}_i$ and the concentration of $\boldsymbol{Z}_i - \boldsymbol{F}_i$, we simply need to update the argument from Theorem 5.1.

To obtain (7.6), decompose

$$\boldsymbol{Z}_i = (\mathbb{E}_{i-1} \boldsymbol{Y}_i)\boldsymbol{Z}_{i-1} + (\boldsymbol{Y}_i - \mathbb{E}_{i-1} \boldsymbol{Y}_i)\boldsymbol{Z}_{i-1}.$$

Since $\mathbb{E}_{i-1} \boldsymbol{Y}_i$ and $\boldsymbol{Z}_{i-1}$ are both measurable with respect to $\mathcal{F}_{i-1}$ and $\mathbb{E}_{i-1}(\boldsymbol{Y}_i - \mathbb{E}_{i-1} \boldsymbol{Y}_i) = \boldsymbol{0}$, the obvious variant of Proposition 4.3 implies that

$$\begin{aligned} \|\!|\!|\boldsymbol{Z}_i|\!|\!|_{p,q}^2 &\le \|\!|\!|(\mathbb{E}_{i-1} \boldsymbol{Y}_i)\boldsymbol{Z}_{i-1}|\!|\!|_{p,q}^2 + \mathrm{C}_p \|\!|\!|(\boldsymbol{Y}_i - \mathbb{E}_{i-1} \boldsymbol{Y}_i)\boldsymbol{Z}_{i-1}|\!|\!|_{p,q}^2 \\ &\le m_i^2 \|\!|\!|\boldsymbol{Z}_{i-1}|\!|\!|_{p,q} + \mathrm{C}_p m_i^2 \sigma_i^2 \|\!|\!|\boldsymbol{Z}_{i-1}|\!|\!|_{p,q}^2. \end{aligned}$$

The second inequality follows from (4.1). This is the same recurrence we obtain in the proof of Theorem 5.1, relation (5.5). The rest of the argument is the same.

To obtain (7.7), decompose

$$\boldsymbol{Z}_i - \boldsymbol{F}_i = \boldsymbol{Y}_i \boldsymbol{Z}_{i-1} - (\mathbb{E}_{i-1} \boldsymbol{Y}_i)\boldsymbol{F}_{i-1} = (\mathbb{E}_{i-1} \boldsymbol{Y}_i)(\boldsymbol{Z}_{i-1} - \boldsymbol{F}_{i-1}) + (\boldsymbol{Y}_i - \mathbb{E}_{i-1} \boldsymbol{Y}_i)\boldsymbol{Z}_{i-1}.$$

As before, Proposition 4.3 implies that

$$\begin{aligned} \|\!|\!|\boldsymbol{Z}_i - \boldsymbol{F}_i|\!|\!|_{p,q}^2 &\le \|\!|\!|(\mathbb{E}_{i-1} \boldsymbol{Y}_i)(\boldsymbol{Z}_{i-1} - \boldsymbol{F}_{i-1})|\!|\!|_{p,q}^2 + \mathrm{C}_p \|\!|\!|(\boldsymbol{Y}_i - \mathbb{E}_{i-1} \boldsymbol{Y}_i)\boldsymbol{Z}_{i-1}|\!|\!|_{p,q}^2 \\ &\le m_i^2 \|\!|\!|\boldsymbol{Z}_{i-1} - \boldsymbol{F}_{i-1}|\!|\!|_{p,q} + \mathrm{C}_p m_i^2 \sigma_i^2 \|\!|\!|\boldsymbol{Z}_{i-1}|\!|\!|_{p,q}^2. \end{aligned}$$

This is the same recurrence that arose when we established Theorem 5.1, relation (5.6). The balance of the argument is identical. □

7.4. **The Spectral Radius.** Products of matrices are closely related to the evolution of discrete-time linear dynamical systems. In this context, it may be more natural to study the *spectral radius* of the matrix product, rather than its spectral norm. Bounds for the spectral radius follow as corollary of our work, owing to the following classical fact.

Fact 7.5 (Schur)**.** *Let* $\boldsymbol{M} \in \mathbb{M}_d$ *be a square matrix. The spectral radius* $\varrho(\boldsymbol{M})$ *is defined as the maximum absolute value of an eigenvalue of* $\boldsymbol{M}$. *It satisfies the variational principle*

$$\varrho(\boldsymbol{M}) = \inf_{\boldsymbol{S} \in \mathbb{M}_d} \left\| \boldsymbol{S}^{-1} \boldsymbol{M} \boldsymbol{S} \right\|.$$

The infimum takes place over all invertible matrices $\boldsymbol{S}$. *In particular* $\varrho(\boldsymbol{M}) \le \|\boldsymbol{M}\|$.

Let us give an indication of the kinds of results that are possible.

Corollary 7.6 (Expectation Bounds for the Spectral Radius)**.** *Consider an independent sequence* $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n\} \subset \mathbb{M}_d$ *of random matrices, and form the product* $\boldsymbol{Z}_n = \boldsymbol{Y}_n \cdots \boldsymbol{Y}_1$. *Let* $\boldsymbol{S} \in \mathbb{M}_d$ *be a fixed invertible matrix, and assume that*

$$\left\| \boldsymbol{S}^{-1} (\mathbb{E}\, \boldsymbol{Y}_i) \boldsymbol{S} \right\| \le m_i \quad \text{and} \quad \left(\mathbb{E} \left\| \boldsymbol{S}^{-1} (\boldsymbol{Y}_i - \mathbb{E}\, \boldsymbol{Y}_i) \boldsymbol{S} \right\|^2 \right)^{1/2} \le \sigma_i m_i \quad \text{for } i = 1, \dots, n.$$

Let $M = \prod_{i=1}^{n} m_i$ *and* $v = \sum_{i=1}^{n} \sigma_i^2$. *Then*

$$\mathbb{E}\, \varrho(\boldsymbol{Z}_n) \le \exp\left(\sqrt{2v\,(2v \vee \log d)}\right) \cdot M.$$

Proof. Combine Corollary 5.4 and Fact 7.5. □

7.5. **Prospects.** We have developed a collection of nonasymptotic bounds for products of random matrices. These results hold under simple and easily verifiable conditions, and they give accurate predictions about the behavior of some particular instances (e.g., products of iid random perturbations of the identity). The proofs are based on foundational results about the geometry of the Schatten classes, and they can easily be adapted to treat variants of the problems under consideration.

A disappointing feature of our results is that they do not account for interactions between the matrix factors. For example, when $\boldsymbol{Y}_i = \mathbf{I} + \boldsymbol{X}_i/n$ for bounded, independent matrix perturbations $\boldsymbol{X}_i$, we have shown that

$$\log \mathbb{E}\, \|\boldsymbol{Y}_n \cdots \boldsymbol{Y}_1\| \le \frac{1}{n} \sum\nolimits_{i=1}^{n} \|\, \mathbb{E}\, \boldsymbol{X}_i\| + O\left(\sqrt{\frac{\log d}{n}}\right).$$

However, when the matrices $\boldsymbol{X}_i$ commute almost surely, it is easy to show the sharper bound

$$\log \mathbb{E}\, \|\boldsymbol{Y}_n \cdots \boldsymbol{Y}_1\| \le \frac{1}{n} \left\| \sum\nolimits_{i=1}^{n} \mathbb{E}\, \boldsymbol{X}_i \right\| + O\left(\sqrt{\frac{\log d}{n}}\right).$$

The results of Emme and Hubert [13] establish that $\lim_{n\to\infty} \log \mathbb{E}\, \|\boldsymbol{Y}_n \cdots \boldsymbol{Y}_1\| = \lim_{n\to\infty} \left\|\sum_{i=1}^{n} \mathbb{E}\, \boldsymbol{X}_i\right\| / n$. It therefore seems reasonable to conjecture that a refined bound of the latter type exists in more generality. The growth bounds discussed in Remark 5.2 imply a statement of the form

$$\log \mathbb{E}\, \|\boldsymbol{Y}_n \cdots \boldsymbol{Y}_1\| \le \log \frac{1}{n} \left\| \prod\nolimits_{i=1}^{n} \mathbb{E}\, \boldsymbol{X}_i \right\| + \text{error},$$

but the error term is not sharp. This type of bound would echo Tropp's improvements [39] to the Ahlswede–Winter results [1] for a sum of independent random matrices. At present, it is not clear whether this refinement is possible, nor what technical arguments would lead there.

Appendix A. Supplementary Proofs

This appendix collects a few additional arguments. First, we establish the sharp form of the result on subquadratic averages, Proposition 4.3, using an elementary method.

Lemma A.1 (Sharp Subquadratic Averages)**.** *Let* $\boldsymbol{X}, \boldsymbol{Y}$ *be random matrices of the same size that satisfy* $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}] = \mathbf{0}$. *When* $2 \le q \le p$,

$$|||\boldsymbol{X} + \boldsymbol{Y}|||_{p,q}^2 \le |||\boldsymbol{X}|||_{p,q}^2 + \mathrm{C}_p |||\boldsymbol{Y}|||_{p,q}^2,$$

where the optimal constant $\mathrm{C}_p := p - 1$.

Proof. Fix a natural number n, and set $\boldsymbol{Z} = n^{-1}\boldsymbol{Y}$. Inequality (4.4) states that

$$D_1 := |||\boldsymbol{X} + \boldsymbol{Z}|||_{p,q}^2 - |||\boldsymbol{X}|||_{p,q}^2 - 2\mathrm{C}_p |||\boldsymbol{Z}|||_{p,q}^2 \le 0.$$

For a parameter $2 \le k \le n$, Corollary 4.2 and Lyapunov's inequality imply that

$$|||\boldsymbol{X} + k\boldsymbol{Z}|||_{p,q}^2 + |||\boldsymbol{X} + (k-2)\boldsymbol{Z}|||_{p,q}^2 \le 2|||\boldsymbol{X} + (k-1)\boldsymbol{Z}|||_{p,q}^2 + 2\mathrm{C}_p |||\boldsymbol{Z}|||_{p,q}^2.$$

Rearranging the last display, we see that

$$\begin{aligned} D_k &:= |||\boldsymbol{X} + k\boldsymbol{Z}|||_{p,q}^2 - |||\boldsymbol{X} + (k-1)\boldsymbol{Z}|||_{p,q}^2 - 2\mathrm{C}_p k |||\boldsymbol{Z}|||_{p,q}^2 \\ &\le |||\boldsymbol{X} + (k-1)\boldsymbol{Z}|||_{p,q}^2 - |||\boldsymbol{X} + (k-2)\boldsymbol{Z}|||_{p,q}^2 - 2\mathrm{C}_p (k-1) |||\boldsymbol{Z}|||_{p,q}^2 = D_{k-1}. \end{aligned}$$

In particular, $D_k \le D_1 \le 0$. Using a telescoping sum,

$$\begin{aligned}\|\!|\boldsymbol{X}+\boldsymbol{Y}|\!\|_{p,q}^2 - \|\!|\boldsymbol{X}|\!\|_{p,q}^2 &= \sum\nolimits_{k=1}^{n} \left(\|\!|\boldsymbol{X}+k\boldsymbol{Z}|\!\|_{p,q}^2 - \|\!|\boldsymbol{X}+(k-1)\boldsymbol{Z}|\!\|_{p,q}^2 \right) \\ &= \sum\nolimits_{k=1}^{n} \left(D_k + 2\mathrm{C}_p k \|\!|\boldsymbol{Z}|\!\|_{p,q}^2 \right) \le \sum\nolimits_{k=1}^{n} 2\mathrm{C}_p k \|\!|\boldsymbol{Z}|\!\|_{p,q}^2 = \mathrm{C}_p \frac{n+1}{n} \|\!|\boldsymbol{Y}|\!\|_{p,q}^2.\end{aligned}$$

Take the limit as $n \to \infty$ to arrive at the stated result. □

Second, we present a basic numerical inequality for weighted sums of exponentials.

Lemma A.2. *Let $a_1, a_2, \dots, a_n$ be a sequence of real numbers. Then*

$$\sum\nolimits_{i=1}^{n} a_i \exp\left(\sum\nolimits_{k=1}^{i-1} a_k \right) \le \exp\left(\sum\nolimits_{i=1}^{n} a_i \right) - 1.$$

Proof. The elementary inequality $a \le \mathrm{e}^a - 1$, valid for $a \in \mathbb{R}$, implies that

$$a_i \exp\left(\sum\nolimits_{k=1}^{i-1} a_k \right) \le \exp\left(\sum\nolimits_{k=1}^{i} a_k \right) - \exp\left(\sum\nolimits_{k=1}^{i-1} a_k \right).$$

Sum the displayed equation over $i = 1, \dots, n$ to verify the claim. □

References

[1] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3):569–579, 2002.

[2] W. Albar, M. Junge, and M. Zhao. Noncommutative versions of the arithmetic-geometric mean inequality, 2017, 1703.00546.

[3] J. M. Altschuler and P. A. Parrilo. Lyapunov exponent of rank one matrices: Ergodic formula and inapproximability of the optimal distribution, 2019, 1905.07531.

[4] Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, New York, second edition, 2010.

[5] K. Ball, E. A. Carlen, and E. H. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones mathematicae*, 115(1):463–482, 1994.

[6] W. Beckner. Inequalities in Fourier analysis. *Ann. of Math. (2)*, 102(1):159–182, 1975.

[7] Y. Benoist and J.-F. Quint. *Random walks on reductive groups*, volume 62. Springer, Cham, 2016.

[8] M. A. Berger. Central limit theorem for products of random matrices. *Trans. Amer. Math. Soc.*, 285(2):777–803, 1984.

[9] R. Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[10] A. Bonami. étude des coefficients de fourier des fonctions de $l^p(g)$. *Ann. Inst. Fourier (Grenoble)*, 20(2):335–402, 1970.

[11] Collectif. Sharp inequalities for martingales and stochastic integrals. In *Colloque Paul Lévy sur les processus stochastiques*, number 157-158 in Astérisque, pages 75–94. Société mathématique de France, 1988.

[12] S. Dartois and P. J. Forrester. Schwinger-dyson and loop equations for a product of square ginibre random matrices. *Journal of Physics A: Mathematical and Theoretical*, 2020.

[13] J. Emme and P. Hubert. Limit laws for random matrix products, Dec 2017, 1712.03698.

[14] A. Furman. Random walks on groups and random transformations. In *Handbook of dynamical systems, Vol. 1A*, pages 931–1014. North-Holland, Amsterdam, 2002.

[15] H. Furstenberg and H. Kesten. Products of random matrices. *Ann. Math. Statist.*, 31:457–469, 1960.

[16] L. Gross. Existence and uniqueness of physical ground states. *J. Functional Analysis*, 10:52–109, 1972.

[17] M. Gürbüzbalaban, A. Ozdaglar, and P. Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, pages 1–36, 2019.

[18] B. Hanin and M. Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, pages 1–36, 2019.

[19] A. Henriksen and R. Ward. Concentration inequalities for random matrix products, Jul 2019, 1907.05833.

[20] A. Israel, F. Krahmer, and R. Ward. An arithmetic-geometric mean inequality for products of three matrices. *Linear Algebra Appl.*, 488:1–12, 2016.

[21] M. Kieburg. Products of Complex Rectangular and Hermitian Random Matrices, Aug 2019, 1908.09408.

[22] M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. Isoperimetry and processes, Reprint of the 1991 edition.

[23] F. Ledrappier. Some asymptotic properties of random walks on free groups. In *Topics in probability and Lie groups: boundary theory*, volume 28, pages 117–152. Amer. Math. Soc., Providence, RI, 2001.

[24] E. H. Lieb. Convex trace functions and the Wigner-Yanase-Dyson conjecture. *Advances in Math.*, 11:267–288, 1973.
[25] F. Lust-Piquard. Inégalités de Khintchine dans C_p $(1 < p < \infty)$. *C. R. Acad. Sci. Paris Sér. I Math.*, 303(7):289–292, 1986.
[26] A. Naor. On the banach-space-valued azuma inequality and small-set isoperimetry of alon–roichman graphs. *Combinatorics, Probability and Computing*, 21(4):623–634, 2012.
[27] A. Nica and R. Speicher. *Lectures on the combinatorics of free probability*, volume 335 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 2006.
[28] E. Oja. A simplified neuron model as a principal component analyzer. *J. Math. Biol.*, 15(3):267–273, 1982.
[29] R. I. Oliveira. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges, 2009, 0911.0600.
[30] G. Pisier. Martingales with values in uniformly convex spaces. *Israel J. Math.*, 20(3-4):326–350, 1975.
[31] N. R. Rao and A. Edelman. The polynomial method for random matrices. *Found. Comput. Math.*, 8(6):649–702, 2008.
[32] B. Recht and C. Re. Beneath the valley of the noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences, 2012, 1202.4184.
[33] É. Ricard and Q. Xu. A noncommutative martingale convexity inequality. *The Annals of Probability*, 44(2):867–882, 2016.
[34] D. Shlyakhtenko. Random matrices and free probability. In *Random Matrices*, number 26. Amer. Math. Soc., Providence, RI, 2019.
[35] R. Speicher. Lecture notes on "free probability theory", 2019, 1908.08125.
[36] T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *J. Fourier Anal. Appl.*, 15(2):262–278, 2009.
[37] N. Tomczak-Jaegermann. The moduli of smoothness and convexity and the Rademacher averages of trace classes $S_p(1 \leq p < \infty)$. *Studia Math.*, 50:163–182, 1974.
[38] J. A. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.
[39] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.
[40] J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.
[41] J. A. Tropp. The expected norm of a sum of independent random matrices: an elementary approach. In *High dimensional probability VII*, volume 71 of *Progr. Probab.*, pages 173–202. Springer, [Cham], 2016.
[42] J. A. Tropp. Second-order matrix concentration inequalities. *Appl. Comput. Harmon. Anal.*, 44(3):700–736, 2018.
[43] J. N. Tsitsiklis and V. D. Blondel. The Lyapunov exponent and joint spectral radius of pairs of matrices are hard—when not impossible—to compute and to approximate. *Math. Control Signals Systems*, 10(1):31–40, 1997.
[44] A. M. Tulino and S. Verdú. Random matrix theory and wireless communications. *Foundations and Trends in Communications and Information Theory*, 1(1):1–182, 2004.
[45] A. Wilkinson. What are Lyapunov exponents, and why are they interesting? *Bull. Amer. Math. Soc. (N.S.)*, 54(1):79–105, 2017.
[46] G. Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation, 2019, 1902.04760.